Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks

Matthew D. Jankowski,* Christopher S. Henry,† Linda J. Broadbelt,‡ and Vassily Hatzimanikatis§

*Mayo Clinic, Rochester, Minnesota 55905; †Mathematics and Computer Science, Argonne National Laboratory, Argonne, Illinois 60439; ‡Department of Chemical and Biological Engineering, McCormick School of Engineering and Applied Sciences, Northwestern University, Evanston, Illinois 60208; and §Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

ABSTRACT A new, to our knowledge, group contribution method based on the group contribution method of Mavrovouniotis is introduced for estimating the standard Gibbs free energy of formation ($\Delta_f G'^\circ$) and reaction ($\Delta_r G'^\circ$) in biochemical systems. Gibbs free energy contribution values were estimated for 74 distinct molecular substructures and 11 interaction factors using multiple linear regression against a training set of 645 reactions and 224 compounds. The standard error for the fitted values was 1.90 kcal/mol. Cross-validation analysis was utilized to determine the accuracy of the methodology in estimating $\Delta_r G'^\circ$ and $\Delta_f G'^\circ$ for reactions and compounds not included in the training set, and based on the results of the cross-validation, the standard error involved in these estimations is 2.22 kcal/mol. This group contribution method is demonstrated to be capable of estimating $\Delta_r G'^\circ$ and $\Delta_f G'^\circ$ for the majority of the biochemical compounds and reactions found in the *i*JR904 and *i*AF1260 genome-scale metabolic models of *Escherichia coli* and in the Kyoto Encyclopedia of Genes and Genomes and University of Minnesota Biocatalysis and Biodegradation Database. A web-based implementation of this new group contribution method is available free at http://sparta.chem-eng.northwestern.edu/cgi-bin/GCM/WebGCM.cgi.

INTRODUCTION

Thermodynamics is increasingly being applied to improve our understanding of the metabolism of microorganisms, especially in the context of constraint-based analysis of genome-scale models of microorganisms (1–4). Constraints based on the laws of thermodynamics have been applied for the determination of feasible ranges for the rates of biochemical reactions and the concentrations of metabolites (2,5). Methods for quantifying the feasible ranges for the Gibbs free energy change of reaction ($\Delta_r G'$) have been applied to the curation of new metabolic reconstructions (4), the systematic assessment of the degree of reversibility of metabolic reactions (6), and the evaluation of the feasibility of biodegradation reactions (S. D. Finley, L. J. Broadbelt, and V. Hatzimanikatis, unpublished). Numerous methods based on thermodynamic constraints and the laws of thermodynamics have also been applied in the study of the regulatory network of the cell (2,5,8). All these studies require that the standard Gibbs free energy change of reaction ($\Delta_r G'^\circ$) be known so that the degree of thermodynamic favorability of the reactions in these systems can be quantified.

Thermodynamic analysis of metabolism based entirely on experimentally measured $\Delta_r G'^\circ$ data has been restricted to either small-scale systems (8) or small subsections of genome-scale systems (5,6) due to the limited amount of experimental data currently available. For example, in the latest genome-scale model of *Escherichia coli* (4), experimentally measured $\Delta_r G'^\circ$ data are available for only 169 (8.1%) of the 2077 reactions in the model (4). Due to this scarcity of experimentally measured values of $\Delta_r G'^\circ$, methods for its estimation are often applied to fill in the gaps in the experimental data. One of the most prevalent techniques for estimating $\Delta_r G'^\circ$ of biochemical reactions is the group contribution method of Mavrovouniotis (9,10). This method allows the rapid calculation of accurate estimations of $\Delta_r G'^\circ$ and the standard Gibbs free energy of formation ($\Delta_f G'^\circ$) for a wide variety of biological reactions and compounds (1). Unlike the group contribution method of Benson (11), this method is tailored for aqueous organic chemistry taking place at neutral pH involving ionic species.

This group contribution method has been applied to the study of the thermodynamic feasibility of numerous native (12–16) and novel (17,18) metabolic pathways. The method has been utilized to estimate $\Delta_f G'^\circ$ and $\Delta_r G'^\circ$ for the majority of the compounds and reactions contained in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (19) and in an earlier genome-scale model of *E. coli* (1). The method has also enabled the development of thermodynamic metabolic flux analysis, a framework for the genome-scale thermodynamic analysis of metabolism that accounts for the effect of metabolite activity levels on the thermodynamic feasibility of biochemical reactions embedded in a metabolic network (2).

In all these applications, the group contribution method of Mavrovouniotis has been demonstrated to be capable of rapidly producing accurate estimates of $\Delta_f G'^\circ$ and $\Delta_r G'^\circ$ for many of the common metabolites in the central metabolic

Submitted November 1, 2007, and accepted for publication March 10, 2008.

Matthew D. Jankowski and Christopher S. Henry contributed equally to this work.

Address reprint requests to Professor Vassily Hatzimanikatis, EPFL, SB ISIC, LCSB, BCH 3110 (Bât. BCH), CH-1015 Lausanne, Switzerland. Tel.: 41-21-6939870; E-mail: vassily.hatzimanikatis@epfl.ch.

Editor: Costas D. Maranas.

0006-3495/08/08/1487/13 $2.00

doi: 10.1529/biophysj.107.124784

pathways. However, the method could not be used to estimate $\Delta_f G'^\circ$ for molecules involving some sulfur, nitrogen, and halogen substructures commonly found in large, genome-scale metabolic models or in databases of biochemical reactions such as the BioCyc (20), Brenda (21), KEGG (22,23), and University of Minnesota Biocatalysis and Biodegradation Database (UM-BBD) (24). Additionally, $\Delta_f G'^\circ$ estimations calculated using the group contribution method of Mavrovouniotis differ significantly from the literature values for $\Delta_f G'^\circ$ of many phosphorylated compounds (25,26), and $\Delta_r G'^\circ$ estimations differ significantly from experimentally observed $\Delta_r G'^\circ$ values for reactions involving the formation (or destruction) of thioester bonds or the formation (or destruction) of conjugated double bonds. Finally, the method of Mavrovouniotis provides only a limited ability to quantify the uncertainty in the $\Delta_f G'^\circ$ and $\Delta_r G'^\circ$ estimates. Although the initial work by Mavrovouniotis provided 68% and 95% confidence intervals of 3 and 5 kcal/mol, respectively, for the overall uncertainty in all estimated $\Delta_f G'^\circ$, no confidence intervals were provided for the uncertainty in the estimated $\Delta_r G'^\circ$. Additionally, insufficient data were provided for the quantification of the uncertainty in each specific $\Delta_f G'^\circ$ estimate calculated using the method. These limitations result in imprecise predictions of uncertainty in estimated $\Delta_f G'^\circ$ and $\Delta_r G'^\circ$ values.

We introduce here an updated and expanded group contribution method which utilizes a larger and more current training set of $\Delta_r G'^\circ$ and $\Delta_f G'^\circ$ data including new tables of thermodynamic data found in the National Institute of Standards and Technology (NIST) database (27) and in the work by Alberty (25,26), Thauer (28,29), and Dolfing (30,31). Due to the availability of additional data, group contribution energies were fit to a number of molecular substructures involving halogens, sulfur, and nitrogen (Table 1) that were not included in Mavrovouniotis's original work. The addition of these new molecular substructures to the group contribution method enables the estimation of $\Delta_f G'^\circ$ for a wider variety of molecules. The method also includes a set of seven new interaction factors to account for the energy contributions of the various types of conjugated double bonds, thioester bonds, and vicinal chlorine atoms (see Methods). Finally, the uncertainty analysis performed allows the uncertainty of each estimated $\Delta_r G'^\circ$ and $\Delta_f G'^\circ$ to be determined based on the uncertainty in the constituent group contribution energies.

METHODS

Group contribution method

The group contribution method was developed as a means of estimating $\Delta_r G'^\circ$ of a reaction based on the molecular structures of the compounds involved in the reaction (9–11). In group contribution methods, the molecular structure of a single compound is decomposed into a set of smaller molecular substructures based on the hypothesis that $\Delta_r G'^\circ$ and $\Delta_f G'^\circ$ can be estimated using a linear model where each model parameter is associated with one of the constituent molecular substructures (or groups) that combine to form the compound. To estimate $\Delta_f G'^\circ$ of the entire compound, the contributions of each of the groups to this property are summed as follows:

$$\Delta_f G'^\circ_{est} = \sum_{i=1}^{N_{gr}} n_i \Delta_{gr} G'^\circ_i, \quad (1)$$

where $\Delta_f G'^\circ_{est}$ is the estimated $\Delta_f G'^\circ$, $\Delta_{gr} G'^\circ_i$ is the contribution of group i to $\Delta_f G'^\circ_{est}$, n_i is the number of instances of group i in the molecular structure, and N_{gr} is the number of groups for which $\Delta_{gr} G'^\circ_i$ is known (i.e., the total number of groups in our database). Similarly, $\Delta_r G'^\circ$ is estimated by summing the contribution of each structural group created or destroyed during the reaction:

$$\Delta_r G'^\circ_{est} = \sum_{i=1}^{m} \nu_i \left(\sum_{j=1}^{N_{gr}} n_j \Delta_{gr} G'^\circ_j \right), \quad (2)$$

where $\Delta_r G'^\circ_{est}$ is the estimated $\Delta_r G'^\circ$, ν_i is the stoichiometric coefficient of species i in the reaction, and m is the number of species involved in the reaction. The advantage of estimating $\Delta_r G'^\circ$ using Eq. 2 instead of using the estimated formation energies is that any structural groups unchanged during the reaction cancel out of Eq. 2, and this can include groups for which $\Delta_{gr} G'^\circ_i$ is unknown.

Determination of the groups involved in a molecular structure

In keeping with the group contribution scheme developed by Mavrovouniotis, this implementation of the group contribution method involves two different kinds of energy contributions: i), contributions from the structural groups that combine together to form the structure of the molecule (Table 1), and ii), contributions from the interaction factors that account for the effect of the interactions between various structural groups on the $\Delta_f G'^\circ$ of a molecule (Table 2). When calculating $\Delta_f G'^\circ_{est}$ of a compound using the group contribution method, the molecular structure of the compound is first broken down into the set of structural groups that combine to form the compound. $\Delta_f G'^\circ_{est}$ can be calculated for a compound only if every single atom involved in the molecular structure of the compound can be assigned to exactly one structural group for which $\Delta_{gr} G'^\circ_i$ is known.

Some of the larger structural groups included in this group contribution method can be further broken down into smaller structural groups. These larger groups, called ''characteristic groups'', were included in the method because the properties of these groups are significantly different from the summed properties of their smaller constituent structural groups. For example, the $-COO^-$ group could be further broken down into the >C=O group and the $-O^-$ group. However, the $\Delta_{gr} G'^\circ_i$ of the $-COO^-$ group is -83.1 kcal/mol, whereas the sum of the $\Delta_{gr} G'^\circ_i$ values for the >C=O group and the $-O^-$ group is -61.2 kcal/mol. The characteristic groups used in this method were originally developed by Mavrovouniotis based on expert knowledge of biochemistry and goodness of fit of the group contribution model to the available experimental data.

Because these characteristic groups exist, often multiple structural groups can be mapped to the same atoms in the molecular structure of a compound. For example, the carbon in a carboxylic acid functional group can be assigned to either the $-COO^-$ group or the >C=O group. When these cases arise, the atoms should always be assigned to the structural group with the smallest search priority number, which is provided along with the $\Delta_{gr} G'^\circ_i$ values in Table 1 (Fig. 1 *A*). The only exception to this rule concerns the phosphate chains found in molecules such as NAD(H) or ATP. If every phosphate in a phosphate chain is assigned to the structural group with the smallest priority number, then every phosphate that is not a terminal phosphate would be assigned to the $-OPO_2^-$- group. This leads to the assignment of the oxygen bridging two neighboring phosphates to two $-OPO_2^-$- groups, which violates the requirement that every atom be assigned to exactly one group. To avoid this violation, terminal phosphate chains (like the phosphate chain in ATP) involving n phosphorus atoms are always decomposed into one $-OPO_3^-$ group and $(n-1)$ $-OPO_2^-$- groups (Fig. 1 *B*). Similarly, internal phosphate chains (like the phosphate chain in NADH) involving n phosphorus atoms were always decomposed into one $-OPO_3^-$- group and $(n-1)$ $-OPO_2$- groups. An algorithm for automatically breaking down molecular structures into the appropriate

TABLE 1 Structural groups used in group contribution method

Description of molecular substructure	$\Delta_{gr}G'^{\circ}$ kcal/mol	SE_{gr} kcal/mol	Frequency	Search priority
Molecular substructures involving halogens				
−Cl (attached to a primary carbon with no other chlorine atoms attached)*	−11.7	0.481	10	4
−Cl (attached to a secondary carbon with no other chlorine atoms attached)*	−10.2	0.600	7	5
−Cl (attached to a tertiary carbon with no other chlorine atoms attached)*	−7.38	0.422	45	6
−Cl (attached to a primary carbon with one other chlorine atom attached)*	−8.54	0.397	4	2
−Cl (attached to a secondary carbon with one other chlorine atom attached)*	−7.18	0.448	3	3
−Cl (attached to a primary carbon with two other chlorine atoms attached)*	−5.55	0.293	4	1
−Br (attached to an aromatic ring)*	2.50	1.26	3	7
−I (attached to an aromatic ring)*	16.6	1.26	3	8
−F (attached to an aromatic ring)*	−43.0	1.26	3	9
Molecular substructures involving sulfur				
$-S^{1-}$*	12.7	2.85	2	6
−SH	−0.740	0.636	260	5
−S-OH*	32.4	3.42	1	1
$-OSO_3^{1-}$*	−156	0.698	8	4
−S-	8.77	0.740	190	8
−S- (participating in a ring)*	0.720	0.706	73	2
−S-S-	5.69	1.20	16	7
$-S^{+}<$*	21.9	2.05	1	3
Molecular substructures involving phosphorous				
$-O-PO_3^{2-}$	−254	0.159	380	3
$-O-PO_2^{2-}$	−205	0.440	149	4
$-O-PO_2^{1-}-$	−208	0.122	490	6
$-O-PO_2^{1-}-$ (participating in a ring)	−190	0.957	11	2
$-O-PO_2^{1-}-O-$	−234	0.438	48	5
$-CO-OPO_3^{2-}-$	−298	0.239	97	1
Molecular substructures involving nitrogen				
$-NH_3^{+}$	−6.25	0.196	236	12
$-NH_2$	2.04	0.331	223	13
$>NH_2^{+}$	5.95	0.900	5	4
>N-	24.4	1.14	9	16
>N- (participating in two fused rings)	12.4	1.10	18	2
>NH	10.5	0.515	250	5
>NH (participating in a ring)	6.18	0.532	108	6
$>NH^{+}$-	15.5	1.17	3	15
>N- (participating in a ring)	22.1	0.617	777	7
$>N^{+}<$*	61.4	1.94	1	18
=NH	−21.7	1.52	6	11
$=NH_2^{+}$	−22.7	1.34	9	10
$=NH^{+}$- (participating in a ring)*	4.37	1.04	5	8
=N-*	16.1	3.16	1	17
=N- (participating in a ring)	4.17	0.572	41	9
$=N^{+}<$ (double bond and one single bond participating in a ring)	13.5	0.672	721	3
$=N^{+}<$ (participating in two fused rings)*	3.77	1.27	10	1
≡N	−32.1	4.34	4	14
Molecular substructures involving oxygen				
$-O^{1-}$*	−32.8	0.934	7	9
−OH	−41.5	0.126	1117	8
−O-	−23.2	0.408	39	10
−O- (participating in a ring)	−36.6	0.902	195	7
>C=O	−28.4	0.180	734	6
>C=O (participating in a ring)	−30.1	0.292	88	3
−CH=O	−30.4	0.164	204	5
$-COO^{1-}$	−83.1	0.111	455	4
−O-CO-	−75.3	0.422	26	2
−O-CO- (participating in a ring)	−71.0	0.787	18	1

(*Continued*)

TABLE 1 (*Continued*)

Description of molecular substructure	$\Delta_{gr}G'^{\circ}$ kcal/mol	SE_{gr} kcal/mol	Frequency	Search priority
Molecular substructures involving unsaturated carbons				
=CH-	12.8	0.242	198	11
=CH- (participating in a nonaromatic ring)	8.46	0.293	755	8
>C=	15.7	0.394	135	13
$=CH_2$	6.87	0.312	110	12
=CH- (participating in one aromatic ring)	4.93	0.142	64	5
>C= (one single bond and one double bond participating in an aromatic ring)	6.95	0.313	66	6
>C= (two single bonds participating in one nonaromatic ring)	11.7	0.362	58	9
>C= (participating in two fused nonaromatic rings)	16.7	0.891	10	3
>C= (participating in two fused rings: one aromatic and one nonaromatic)	6.77	0.607	9	4
>C= (double bond and one single bond participating in a ring)	32.1	2.14	3	7
>C= (participating in two fused aromatic rings)	−0.0245	0.927	4	2
≡CH	60.7	4.74	1	10
≡C-	41.6	2.32	3	1
Molecular substructures involving saturated carbons				
$-CH_3$	−3.65	0.109	332	6
$>CH_2$	1.62	0.0880	916	7
$>CH_2$ (participating in one ring)	3.18	0.247	781	3
>CH-	5.08	0.153	981	8
>CH- (participating in one ring)	4.84	0.216	409	4
>CH- (participating in two fused rings)	2.60	0.779	30	1
>C<	7.12	0.298	148	9
>C< (participating in one ring)	7.17	0.420	153	5
>C< (participating in two fused rings)*	−3.89	3.03	1	2

*These groups were not part of the group contribution method of Mavrovouniotis.

structural groups in the group contribution method is discussed in the work by Forsythe, Karp and Mavrovouniotis (32).

The molecular structures being decomposed into structural groups must also be in the form of the predominant ion for the molecule in the same conditions at which the fitting of the $\Delta_{gr}G_i'^{\circ}$ values was performed: pH 7, zero ionic strength, and a temperature of 298 K. The predominant ions of all the molecules involved in the training set at pH 7 were determined using pK_a estimation software (MarvinBeans pK_a estimation plug-in, ver. 4.0.3, ChemAxon, Budapest, Hungary). When a molecule exists in multiple isomeric or resonance forms in equilibrium, such as keto-enol tautomers, the most stable form (the form resulting in the lowest $\Delta_f G'^{\circ}_{est}$) is decomposed into structural groups. This ensures that the form of the molecule used to calculate $\Delta_r G'^{\circ}_{est}$ is the predominant form in solution.

Stereochemistry is ignored when labeling atoms in a molecule according to their structural groups. For example, all forms of sugars with six carbon atoms including glucose, galactose, and mannose, which have $\Delta_f G'^{\circ}_{obs}$ values of −219, −217, and −217 kcal/mol, respectively, are decomposed into exactly the same structural groups and interaction factors, and as a result, all these sugars have identical $\Delta_f G'^{\circ}_{est}$ values. This is a reasonable assumption given the similarity of the $\Delta_f G'^{\circ}_{obs}$ values.

TABLE 2 Interaction factors used in the group contribution method

Interaction factor	$\Delta_{gr}G'^{\circ}$ kcal/mol	SE_{gr} kcal/mol	*t*-test	Frequency	SE_{MLR} with/without
Heteroaromatic rings	−1.95	0.339	0.00	736	1.90/1.91
Three-member rings	14.4	1.56	0.00	2	1.90/1.92
Hydrocarbon	3.68	0.865	0.00	7	1.90/1.91
Amide	−14.3	0.348	0.00	122	1.90/2.33
Thioester*	−11.3	0.459	0.00	161	1.90/2.07
Vicinal Cl*	1.92	0.356	0.00	35	1.90/1.91
OCCC conjugation*	−1.55	0.240	0.00	365	1.90/1.91
OCCO conjugation*	2.46	0.183	0.00	374	1.90/1.95
OCCN conjugation*	−3.02	0.582	0.00	14	1.90/1.91
CCCN conjugation*	−5.29	0.717	0.00	15	1.90/1.92
CCCC conjugation*	−4.82	0.419	0.00	44	1.90/1.94

*These groups were not part of the group contribution method of Mavrovouniotis.

Once every single atom in the molecular structure of a compound has been assigned to the proper structural group, the interaction factors must be determined. The $\Delta_{gr}G_i'^{\circ}$ associated with each interaction factor is then added to compound $\Delta_f G'^{\circ}_{est}$ to account for the effect of the interaction factor on the formation energy. There are seven types of interaction factors used in this implementation of the group contribution method (Fig. 1 *C*). Four of the interaction factors used were originally proposed in the group contribution method of Mavrovouniotis: the hydrocarbon factor, the heteroaromatic ring factor, the three-member ring factor, and the amide factor. The hydrocarbon factor is added to $\Delta_f G'^{\circ}_{est}$ of any compound that consists of only carbon and hydrogen. The heteroaromatic ring factor is added to $\Delta_f G'^{\circ}_{est}$ of a compound for every heteroaromatic ring in the compound, as determined according to Hückel's rule. Similarly, the three-member ring factor is added to $\Delta_f G'^{\circ}_{est}$ of a compound for every three-member ring in the compound regardless of the atoms that make up the ring. The amide factor is added to $\Delta_f G'^{\circ}_{est}$ of a compound for every instance of a nitrogen atom neighboring a carbonyl group in the compound. Note that if a nitrogen atom is neighboring two carbonyl groups, this is counted as a single amide factor.

Three new types of interaction factors were introduced in this implementation of the group contribution method that were not included in the method of Mavrovouniotis: the thioester factor, the double bond conjugation factors, and the vicinal Cl factor. The thioester factor is added to $\Delta_f G'^{\circ}_{est}$ of a compound for every instance of a sulfur atom neighboring a carbonyl group in the compound. This factor accounts for high energy of the thioester bond (33). Like the amide factor, if a sulfur atom is neighboring two carbonyl groups, this is counted as a single thioester factor.

The conjugation of double bonds has a significant stabilizing effect on the molecular structure of a molecule, making the removal of a conjugated double bond more difficult than the removal of an isolated double bond (33).

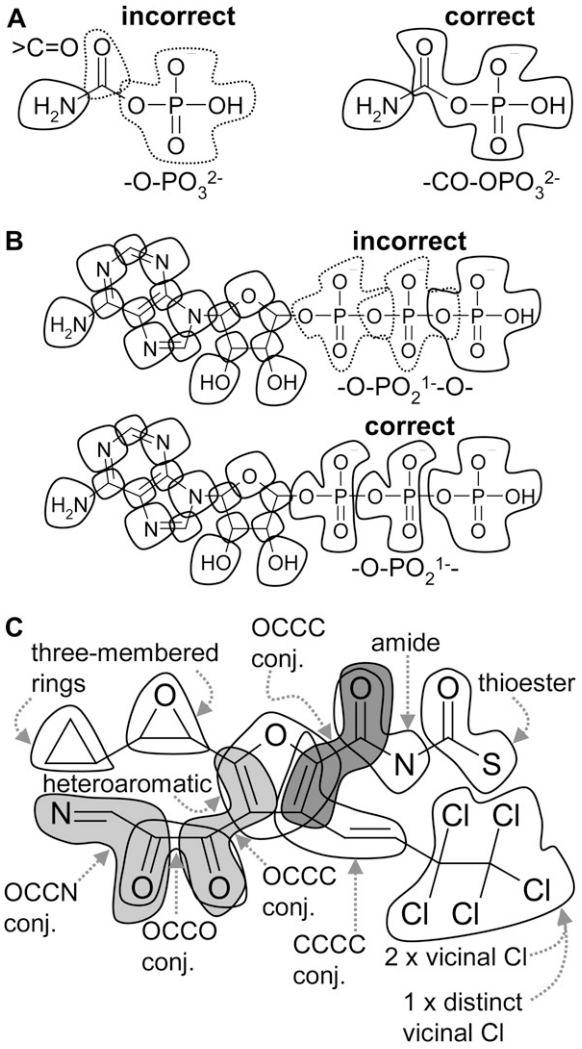


FIGURE 1 Decomposition of molecular structures into structural groups and interaction factors. When assigning atoms in a molecular structure to structural groups, atoms should always be assigned to the structural group with the lowest search priority number (*A*). Phosphate chains such as ATP and NADH are the only exceptions to this rule; a phosphate chain of size n should be decomposed into $(n-1)$ -O-PO_2^{1-}- groups and one $-$O-PO_2^{1-}-O- or $-$O-PO_3^{1-} group (*B*). Although each atom in the molecular structure may only participate in a single structural group, atoms can participate in multiple interaction factors. All but one of the interaction factors included in this group contribution method are found within the structure of the example molecule in (*C*) (the hydrocarbon factor is not included). Note that conjugated double bonds contained entirely within an aromatic or heteroaromatic ring are not counted. However, double bonds outside the ring conjugated to double bonds within the ring are counted. Also note that nitrogen atoms neighboring two carbonyl groups are only counted as a single amide.

Without any interaction factor for double bond conjugation, the group contribution method has no means of capturing these characteristics of conjugated double bonds. Therefore, the double bond conjugation factors were introduced to account for the stabilizing effect of double bond conjugation on a molecular structure. Ten forms of conjugated double bonds are possible in a molecular structure containing C, N, and O, and a separate double bond conjugation factor was initially introduced for each of these 10 forms (Table 3). Five of these forms were not included in the final implementation of the method due to a lack of data or because the conjugation factor was statistically insignificant (see Table 3 and Results). Note that double bond conjugation factors are not added for conjugated double bonds that are contained completely within an aromatic or heteroaromatic ring.

The vicinal Cl factor was introduced based on the examination of the effect of chlorine substitution on the $\Delta_f G'^{\circ}_{obs}$ of aliphatic compounds performed by Dolfing and Janssen (31). Dolfing and Janssen proposed that chlorine atoms attached to neighboring carbon atoms have a destabilizing effect on one another, and an interaction factor is required to account for this destabilization to accurately estimate the $\Delta_f G'^{\circ}_{est}$ of chlorinated compounds using the group contribution method. The vicinal Cl factor is an implementation of the interaction factor proposed by Dolfing and Janssen, and two variations of this interaction factor were explored. The first variation implemented, $VCl_{distinct}$, is based on the hypothesis that a larger number of chlorine atoms attached to neighboring carbons results in a larger destabilizing effect, described mathematically as follows:

$$VCl_{distinct} = \Delta_{gr}G'^{\circ}_{VCl_{distinct}}\left(\sum_{i=1}^{N_C-1}\sum_{j=i+1}^{N_C}\delta_{ij}\min(N_{Cl,i}, N_{Cl,j})\right), \quad (3)$$

where $VCl_{distinct}$ is the total value of the correction for the interaction of vicinal chlorine atoms that is added to $\Delta_f G'^{\circ}_{est}$, $\Delta_{gr}G'^{\circ}_{VCl_{distinct}}$ is the group contribution energy for the vicinal Cl interaction factor, N_C is the number of carbon atoms in the molecule, $N_{Cl,i}$ is the number of chlorine atoms attached to carbon atom i, and δ_{ij} is the Kronecker Δ, a binary variable equaling zero unless carbon atom i is bonded to carbon atom j.

The second variation of the vicinal Cl interaction factor, VCl_{binary}, is based on the hypothesis that the destabilizing effect of vicinal chlorine atoms is independent of the number of chlorine atoms attached to each of the neighboring carbons (Fig. 1 *C*), described mathematically as follows:

$$VCl_{binary} = \Delta_{gr}G'^{\circ}_{VCl_{binary}}\left(\sum_{i=1}^{N_C-1}\sum_{j=i+1}^{N_C}\delta_{ij}\right). \quad (4)$$

Both variations of the vicinal Cl interaction factor were tested, and the $VCl_{distinct}$ interaction was selected for the final implementation of the method because it resulted in the best possible fit of the thermodynamic data included in the training set (see Results).

Multiple linear regression

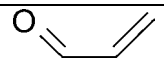
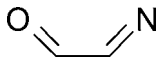
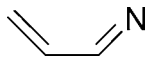
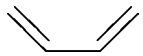
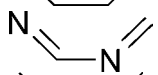
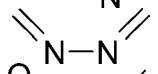
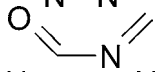
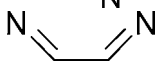
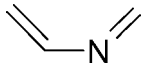
The multiple linear regression (MLR) method (least squares) was used to determine the $\Delta_{gr}G'^{\circ}_i$ values for the set of structural groups and interaction factors that allow the best fit of the observed $\Delta_r G'^{\circ}$ ($\Delta_r G'^{\circ}_{obs}$) and observed $\Delta_f G'^{\circ}$ ($\Delta_f G'^{\circ}_{obs}$) values included in a training set. The $\Delta_{gr}G'^{\circ}_i$ values are calculated using the following:

$$\Delta_{gr}\mathbf{G}'^{\circ} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Delta\mathbf{G}'^{\circ}_{obs}), \quad (5)$$

where $\mathbf{\Delta}_{gr}\mathbf{G}'^{\circ}$ is an $N_{gr} \times 1$ vector of the energies associated with each group in the group contribution method, $\mathbf{X}$ is an $N_{obs} \times N_{gr}$ matrix of the number of each group contained in each molecular structure or created or destroyed in each reaction in the training set, $\mathbf{X}'$ is the transpose of matrix $\mathbf{X}$, N_{obs} is the number of $\Delta_r G'^{\circ}_{obs}$ and $\Delta_f G'^{\circ}_{obs}$ values included in the training set, and $\mathbf{\Delta G}'^{\circ}_{obs}$ is an $N_{obs} \times 1$ vector of $\Delta_r G'^{\circ}_{obs}$ and $\Delta_f G'^{\circ}_{obs}$ values included in the training set (34).

MLR is the ideal technique for producing $\Delta_{gr}\mathbf{G}'^{\circ}_i$ values that optimally fit the training set only if the data included in the training set satisfies the following two conditions: (MLR.I) $\Delta_r G'^{\circ}$ and $\Delta_f G'^{\circ}$ must be linearly related to the model parameters (the $\Delta_{gr}G'^{\circ}_i$ values) and the differences between the $\Delta G'^{\circ}_{obs,i}$ and $\Delta G'^{\circ}_{est,i}$ for each data point in the training set must be uncorrelated, and (MLR.II) the absolute uncertainty in each of the $\Delta_r G'^{\circ}$ and $\Delta_f G'^{\circ}$ observations included in $\mathbf{\Delta G}'^{\circ}_{obs}$ must be similar in magnitude (34). The random distribution of the residuals of the fit indicates that condition MLR.I is satisfied (see Fig. 2 *B*). The discussion of the uncertainty in the training set data explains why condition MLR.II is also satisfied by the data (see the section ''Uncertainty in training set data'').

TABLE 3 Interaction factors for conjugated double bonds

Name	Image	$\Delta_{gr}G'^{\circ}$ kcal/mol	SE_{gr} kcal/mol	*t*-test	Frequency	SE_{MLR} with/without
OCCC conjugation		−1.55	0.233	0.00	372	1.90/1.91
OCCO conjugation		2.46	0.180	0.00	381	1.90/1.95
OCCN conjugation		−3.02	0.578	0.00	14	1.90/1.91
CCCN conjugation		−5.29	0.710	0.00	15	1.90/1.92
CCCC conjugation		−4.82	0.412	0.00	44	1.90/1.94
NCNC conjugation*		None	None	None	0	None
CNNC conjugation*		None	None	None	0	None
OCNC conjugation†		−1.92	0.821	0.02	16	1.90/1.90
NCCN conjugation†		−0.746	0.902	0.41	5	1.90/1.90
CCNC conjugation†		0.530	0.569	0.35	39	1.90/1.90

*Removed from method due to lack of data.
†Removed from method due to high *t*-test.

Quantification of the goodness of fit

The goodness of the MLR fit is quantified using the standard deviation of the differences between $\mathbf{\Delta G'^{\circ}_{obs}}$ and $\mathbf{\Delta G'^{\circ}_{est}}$ for the compounds and reactions involved in the training set, SE_{MLR} (34):

$$SE_{MLR} = \sqrt{\mathbf{R}'\mathbf{R}/(N_{obs} - N_{gr})}, \quad (6)$$

where $\mathbf{R}'$ is the transpose of the vector $\mathbf{R}$, and $\mathbf{R}$ is the vector of residuals for the fit, calculated as follows (34):

$$\mathbf{R} = \mathbf{\Delta G'^{\circ}_{obs}} - \mathbf{X\Delta_{gr}G'^{\circ}}. \quad (7)$$

If the differences between $\mathbf{\Delta G'^{\circ}_{obs}}$ and $\mathbf{\Delta G'^{\circ}_{est}}$ in the training set follow a normal distribution, then 68% of the residuals will be less than or equal to SE_{MLR} (34). The SE_{MLR} is also used to assess the effect of removal or addition of interaction factors on the group contribution scheme (see the section ''Whole model and individual parameter validation'').

Formation of the training set for the MLR

The $\mathbf{\Delta G'^{\circ}_{obs}}$ values used in the training set for the MLR involved a total of 3153 $\Delta_r G'^{\circ}_{obs}$ values and 288 $\Delta_f G'^{\circ}_{obs}$ values. The $\Delta_r G'^{\circ}_{obs}$ and $\Delta_f G'^{\circ}_{obs}$ values in the training set were pulled from a variety of literature sources including work on methanogenesis by Thauer (28,29), work on halogen thermodynamics by Dolfing and co-workers (30,31), work on formation energy standardization and redox potentials by Alberty (25,26), and thermodynamic data compiled in the NIST (27) and National Bureau of Standards (NBS) (35) databases. The experimentally measured $\Delta_r G'^{\circ}_{obs}$ values reported in these references were captured under a variety of temperature and pH conditions. Only data captured within one pH unit and 15 K of the chosen reference state of pH 7 and 298 K was utilized. Most of the data utilized were collected within 1 K of 298 K and 0.1 pH units of pH 7 (Fig. 3). Overall, 645 distinct biochemical reactions are represented in the 3153 $\Delta_r G'^{\circ}_{obs}$ values used in the training set, meaning that multiple data points were included for many reactions. Similarly, 224 distinct molecular structures are represented by the 288 $\Delta_f G_{obs}{}'^{\circ}$ values used in the training set. When multiple data points existed for single reactions or compounds, we used all data points in the data set rather than averaging the data and including the average. By using all the data points instead of the average, the variability in the data is included in the residuals, covariance matrix, and standard deviation for the fit, which results in a better quantification of the uncertainty in the group free energy values. All $\Delta_r G'^{\circ}_{obs}$ and $\Delta_f G'^{\circ}_{obs}$ values included in the training set are listed in Supplementary Material, Data S2 along with the associated reactions and compounds.

Uncertainty in training set data

To estimate the total uncertainty in each $\Delta_f G'^{\circ}_{obs}$ and $\Delta_r G'^{\circ}_{obs}$ data point included in the training set, the sources of uncertainty were enumerated and quantified. The total uncertainty in the $\Delta_f G'^{\circ}_{obs}$ values included in the training set were estimated from the precision of the reported $\Delta_f G'^{\circ}_{obs}$ values; the reported precision in the $\Delta_f G'^{\circ}_{obs}$ values ranges from 0.01 to 1 kcal/mol, implying that the absolute uncertainty in the $\Delta_f G'^{\circ}_{obs}$ values ranges from 0.005 to 0.5 kcal/mol (26,28,35,36). The total uncertainties in the $\Delta_r G'^{\circ}_{obs}$ values included in the training set were calculated from four primary sources: (UC.I) uncertainty in the method used to measure the equilibrium constant, (UC.II) uncertainty due to differences between the ionic strength at which each $\Delta_r G'^{\circ}_{obs,i}$ was measured and the reference ionic strength of zero, (UC.III) uncertainty due to differences between the pH at which each $\Delta_r G'^{\circ}_{obs,i}$ was measured and the reference pH of 7, and (UC.IV) uncertainty due to differences between the temperature at which each $\Delta_r G'^{\circ}_{obs,i}$ was measured and the reference temperature of 298 K.

Most of the $\Delta_r G'^{\circ}_{obs,i}$ values included in the training set were measured using spectroscopy, which has a typical precision of 1%–3% of the measured values when used to determine equilibrium constants (37). This translates into an absolute uncertainty of <0.30 kcal/mol for 95% of the $\Delta_r G'^{\circ}_{obs}$ values included in the training set. Uncertainty due to deviations of the conditions for the $\Delta_r G'^{\circ}_{obs}$ measurements from the reference ionic strength of zero was determined using the extended Debye-Huckel equation as described in Maskow and Stockar (12). For 95% of the reactions in the training set, this uncertainty was <1.45 kcal/mol when the deviation in ionic strength was <0.2 M. The absolute uncertainty due to deviations in the conditions for the

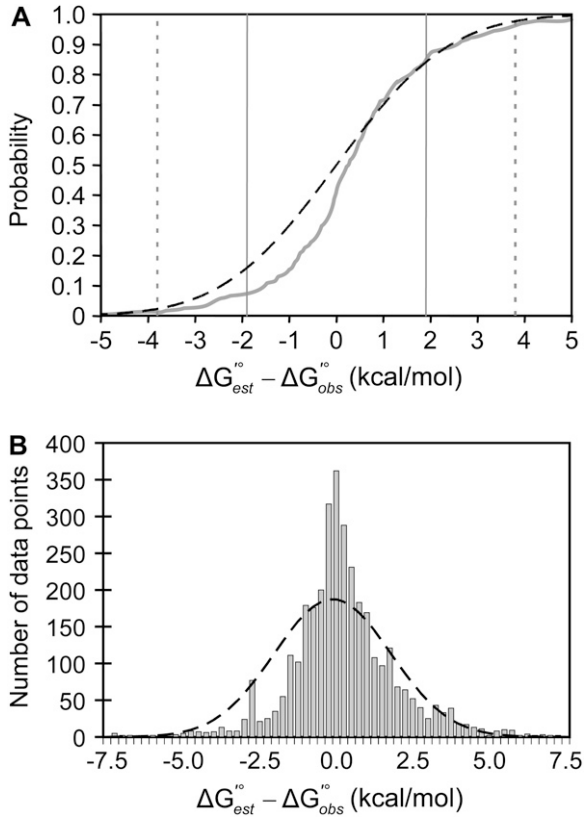


FIGURE 2 Distribution of residuals from the MLR fitting of the training set cumulative distribution (*A*) and histogram (*B*) of the deviations between $\Delta G'^{\circ}_{\text{est}}$ calculated using the fitted $\Delta_{\text{gr}}G'^{\circ}$ values and the $\Delta G'^{\circ}_{\text{obs}}$ values in the training set. The cumulative probability for the deviations between $\Delta G'^{\circ}_{\text{est}}$ and $\Delta G'^{\circ}_{\text{obs}}$ (*solid gray line* in *A*) nearly overlaps with the cumulative probability for a normal distribution (*dashed line* in *A*). The points of intersection between the cumulative probability line for the residuals of the fitting with the SE_{MLR} lines (*solid vertical gray lines*) and 2 SE_{MLR} lines (*dashed vertical gray lines*) indicate that ~85% and 96% of the $\Delta G'^{\circ}_{\text{est}}$ values will fall within one and two standard deviations, respectively, of $\Delta G'^{\circ}_{\text{obs}}$. The distribution of deviations (*shaded bars* in *B*) between $\Delta G'^{\circ}_{\text{est}}$ and $\Delta G'^{\circ}_{\text{obs}}$ is more compact than a normal distribution (*dashed line* in *B*) with the same standard deviation (1.90 kcal/mol). This confirms that uncertainty estimations based on standard deviations will be more conservative than expected for normally distributed errors.

$\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ measurements from the reference pH of 7 was <1.49 kcal/mol for 95% of the training set reactions within the allowed pH ranges (pH 6–8), as calculated using the methods described by Alberty (26). Uncertainties due to pH and ionic strength deviations are independent of the reference $\Delta G'^{\circ}_{\text{obs}}$ value.

As $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ measurements were accepted into the training set if measured within 15 K of the reference temperature of 298 K, deviations of the $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ measurement conditions from the reference temperature were another source of uncertainty in the $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ values. A rearranged version of the Gibbs-Helmholtz relationship was utilized to determine how temperature changes affect $\Delta_{\text{r}}G'^{\circ}$ of a reaction:

$$\frac{T\partial\Delta_{\text{r}}G'^{\circ}}{\Delta_{\text{r}}G'^{\circ}\partial T} = 1 - \frac{\Delta_{\text{r}}H'^{\circ}}{\Delta_{\text{r}}G'^{\circ}}, \tag{8}$$

where $\Delta_{\text{r}}H'^{\circ}$ is the standard enthalpy change of reaction. Although measured $\Delta_{\text{r}}H'^{\circ}$ values are unavailable for most of the reactions contained in the training set, the $(1 - \Delta_{\text{r}}H'^{\circ}/\Delta_{\text{r}}G'^{\circ})$ term in the Gibbs-Helmholtz relationship will typically have a maximum value of one for biochemical reactions. Based on this assumption, a 15 K maximum temperature change results in a maximum change of 5.7% in $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ This translates into an absolute error of <0.57 kcal/mol for 95% of the $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ values in the training set. Overall, for 95% of the $\Delta G'^{\circ}_{\text{obs}}$ values included in the training set, the total absolute uncertainty is >0.1 kcal/mol and <2.2 kcal/mol, which satisfies the MLR.II condition that the uncertainty of all $\Delta G'^{\circ}_{\text{obs}}$ values in the training set be similar in magnitude.

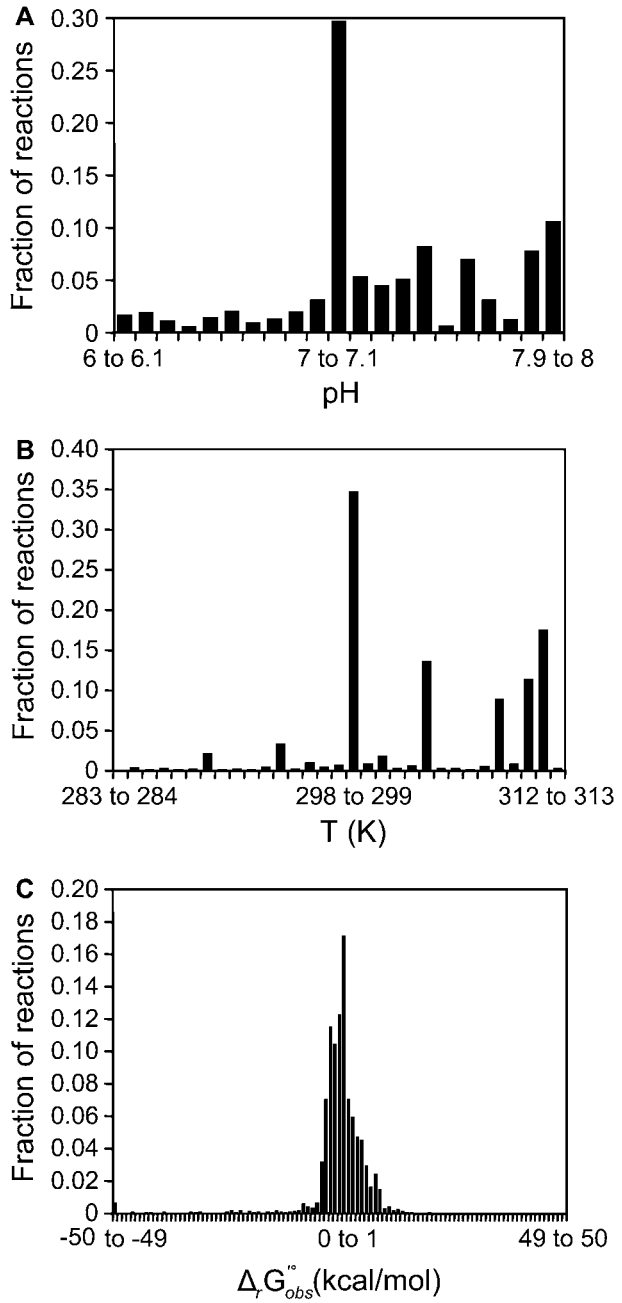


FIGURE 3 pH, temperature, and $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ distributions for the $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ data within the training set. The distributions of pH (*A*), *T* (*B*), and $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ (*C*) values for the 3153 $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ measurements used in the training set to determine the group contribution energies are shown. The most prevalent condition for the $\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ measurements included in the training set was pH 7.0–7.1 and 298–299 K, which is the reference state selected for this group contribution method. Interestingly, most of the$\Delta_{\text{r}}G'^{\circ}_{\text{obs}}$ values used in the fitting have an absolute value of <10 kcal/mol.

Quantification of the uncertainty in the $\Delta_{gr}G'^{\circ}$ values

The uncertainties in the $\Delta_{gr}G'^{\circ}_i$ values estimated using MLR were quantified using the covariance matrix of the MLR, which allowed the calculation of a standard error for each $\Delta_{gr}G'^{\circ}_i$ in the group contribution method as follows (34):

$$SE_{gr,i} = \sqrt{(SE^2_{MLR}(\mathbf{X}'\mathbf{X})^{-1})_{i,i}}, \quad (9)$$

where $SE_{gr,i}$ is the standard error for the group contribution value of group i, $\Delta_{gr}G'^{\circ}_i$. The $SE_{gr,i}$ values can be used to quantify the uncertainty in the estimated Gibbs free energy of formation and reaction, $\Delta G'^{\circ}_{est}$, calculated by taking the Euclidean norm of the uncertainties in each group $\Delta_{gr}G'^{\circ}_i$ value multiplied by the number of instances of each group involved in the molecular structure or reaction (34):

$$SE_{\Delta G'^{\circ}_{est,j}} = \sqrt{\sum_{i=1}^{N_{gr}} (X_{i,j}SE_{gr,i})^2}. \quad (10)$$

Whole model and individual parameter validation

An F-test was performed to validate the use of the linear group contribution model to estimate $\Delta_rG'^{\circ}$ and $\Delta_fG'^{\circ}$ for the data included in the training set. The F-test indicates whether or not the variability in the $\Delta G'^{\circ}_{obs}$ values within the training set that is captured by the group contribution model is statistically significant compared to the variability not captured by the model (the variances between $\Delta G'^{\circ}_{obs}$ and $\Delta G'^{\circ}_{est}$) (34). If the location of the F-value in the F-cumulative distribution function corresponds to a probability value >90%, the linear model is accepted.

A *t*-test was also used to validate the inclusion of each interaction factor in the group contribution model. The *t*-test indicates whether the value of $\Delta_{gr}G'^{\circ}_i$ for an interaction factor is statistically significant compared to the uncertainty in the $\Delta_{gr}G'^{\circ}_i$ value, $SE_{gr,i}$ (34). The interaction factor was retained as a part of the model if the location of its *t* value in the student t-cumulative distribution function corresponds to a probability value of <5% (34). Although *t*-tests were performed on the structural groups as well, structural groups with high *t*-tests were not removed from the model because they were required for the complete decomposition of the molecular structures involved in the training set. For example, although the >C= group that participates in two fused aromatic rings has a $\Delta_{gr}G'^{\circ}_i$ of −0.0245 kcal/mol and an $SE_{gr,i}$ of 0.927 kcal/mol resulting in a *t*-test of 0.98, it is retained because it is required for the complete decomposition of molecules involving fused aromatic rings.

Interaction factors with *t*-tests of over 5% (indicating insignificantly small $\Delta_{gr}G'^{\circ}_i$ values) were eliminated from the final implementation of the group contribution method because interaction factors are not required for the complete decomposition of a molecular structure and removal of an interaction factor with a $\Delta_{gr}G'^{\circ}_i$ of zero results in little or no increase in SE_{MLR} of the fitting. However, passing the *t*-test does not guarantee that the addition of an interaction factor results in any significant reduction in SE_{MLR} for the fitting. Therefore, in addition to performing a *t*-test for each interaction factor, the SE_{MLR} with and without the interaction factor was also calculated as a measure of the effect of the interaction factor on the goodness of fit. Details of how the *t*-tests and F-test were calculated are provided in Data S1.

Cross-validation analysis

A cross-validation analysis of the training set used for the fitting was performed to validate the ability of the group contribution method to produce $\Delta_fG'^{\circ}_{est}$ and $\Delta_rG'^{\circ}_{est}$ estimates for compounds and reactions outside the training set with the same degree of accuracy as the $\Delta_fG'^{\circ}_{est}$ and $\Delta_rG'^{\circ}_{est}$ estimates for compounds and reactions within the training set. Two hundred distinct cross-validation runs were performed. In each run, 10% of the 869 distinct reactions and compounds involved in the training set were selected at random, and all the $\Delta G'^{\circ}_{obs}$ values associated with each of the selected compounds and reactions were removed from the training set. When a compound was removed from the training set, the $\Delta G'^{\circ}_{obs}$ values associated with the stereoisomeric forms of the compound were also removed from the data set. However, reactions involving the removed compounds were left in the training set unless they were also randomly selected for removal. MLR was then performed on the data remaining in the training set to produce a new set of $\Delta_{gr}G'^{\circ}_i$ values. The SE_{MLR}, $\mathbf{\Delta G'^{\circ}_{est}}$, $\mathbf{SE}_{\Delta G'^{\circ}_{est}}$, and $\mathbf{R}$ were all calculated for the data included and excluded from the reduced training set using the new set of $\Delta_{gr}G'^{\circ}_i$ values.

RESULTS

Development of the improved group contribution method

The new, to our knowledge, group contribution method introduced here consists of 74 molecular substructures (called structural groups) and 11 factors to account for interactions between molecular substructures (called interaction factors) for which group contribution energies ($\Delta_{gr}G'^{\circ}_i$) are provided (Tables 1 and 2). The $\Delta_{gr}G'^{\circ}_i$ values provided were determined based on an MLR fitting of a training set consisting of 224 compounds with 288 known $\Delta_fG'^{\circ}$ values and 645 reactions with 3153 known $\Delta_rG'^{\circ}$ values. The standard error for the fit of the group contribution model to this training set was 1.90 kcal/mol.

Although this new group contribution method is based on the previous group contribution method developed by Mavrovouniotis (9,10), the new method is a significant improvement over the previous method both in the range of biochemical compounds and reactions for which $\Delta_fG'^{\circ}$ and $\Delta_rG'^{\circ}$ may be estimated and in the accuracy of the $\Delta_fG'^{\circ}$ and $\Delta_rG'^{\circ}$ estimates generated. The expanded applicability of this new group contribution method is due to the addition of 20 new structural groups to the method. When restricted to the structural groups included in the previous group contribution method, $\Delta_fG'^{\circ}$ could be estimated for only 65% of the compounds and $\Delta_rG'^{\circ}$ could be estimated for only 97% of the reactions in the training set for the new method. In contrast, the new method allows the estimation of $\Delta_fG'^{\circ}$ and $\Delta_rG'^{\circ}$ for 100% of the compounds and reactions in the training set.

The expanded applicability of the new group contribution method also extends to large databases of known biochemical reactions such as the KEGG, UM-BBD, *i*AF1260 (4), and *i*JR904 (38). The application of the current and previous group contribution methods to the estimation of $\Delta_fG'^{\circ}$ and $\Delta_rG'^{\circ}$ for these databases is discussed in detail later (see the section ''Estimating $\Delta G'^{\circ}$ of known biochemical reactions''). For the compounds and reactions in the training set for which $\Delta_fG'^{\circ}$ and $\Delta_rG'^{\circ}$ could be estimated using the previous group contribution method, the standard error of the estimates generated by the previous method was 3.92 kcal/mol, compared to a standard error of 1.98 kcal/mol for the estimates generated by the new group contribution method.

This difference in standard error confirms that the accuracy in the $\Delta G'^{\circ}$ estimates produced using the new group contribution method is significantly improved.

Results from MLR fitting

To assess the goodness of fit of the new group contribution method to the training set of available thermodynamic data, the distribution of the residuals of the fit (the deviations between the estimated $\Delta G'^{\circ}$ ($\Delta G'^{\circ}_{\mathrm{est}}$) and the observed $\Delta G'^{\circ}$ ($\Delta G'^{\circ}_{\mathrm{obs}}$) values) were analyzed (Fig. 2). Analysis of the cumulative distribution of the residuals indicated that 85% and 96% of $\Delta G'^{\circ}_{\mathrm{est}}$ for the training set fall within one and two standard deviations of $\Delta G'^{\circ}_{\mathrm{obs}}$, respectively (Fig. 2 *A*). This agrees well with the confidence intervals expected if the residuals from the training set were normally distributed (68% and 95% within one and two standard deviations, respectively). The distribution of residuals for the training set (*shaded bars* in Fig. 2 *B*) is also similar to a normal distribution with the same mean and standard deviation (*dashed line* in Fig. 2 *B*). The high peak in the distribution of the residuals above the expected normal distribution indicates the presence of a small number of outlying data points with uncharacteristically large errors that are causing the standard deviation to be larger than would be expected. Although the reactions and compounds associated with each of these outlying data points were carefully analyzed, no clear trends emerged to indicate the need for any additional structural groups or interaction factors in the group contribution method.

F-tests and *t*-tests were also performed to validate the group contribution method as a whole and to validate each of the interaction factors included in the group contribution method (see Methods). The F-value calculated for the method corresponded to a probability of 100% on the F-cumulative distribution curve, indicating that the method passes the F-test. Additionally, all the *t*-tests for the interaction factors included in the final implementation of the new method scored below 5%, indicating that the $\Delta_{\mathrm{gr}}G'^{\circ}_{\mathrm{i}}$ values for these factors are statistically significant.

The uncertainties in the $\Delta_{\mathrm{gr}}G'^{\circ}_{\mathrm{i}}$ values of the structural groups (Table 1) and interaction factors (Table 2) ($SE_{\mathrm{gr,i}}$) were utilized to calculate the specific uncertainty in $\Delta_{\mathrm{f}}G'^{\circ}_{\mathrm{est}}$ or $\Delta_{\mathrm{r}}G'^{\circ}_{\mathrm{est}}$ ($SE_{\Delta G'^{\circ}_{\mathrm{est}}}$) for each data point in the training set (see Methods). We found that 73% and 87% of the $\Delta G'^{\circ}_{\mathrm{est}}$ values in the training set fell within one and two $SE_{\Delta G'^{\circ}_{\mathrm{est}}}$ of the $\Delta G'^{\circ}_{\mathrm{obs}}$ values, respectively, validating that the $SE_{\Delta G'^{\circ}_{\mathrm{est}}}$ calculated from the $SE_{\mathrm{gr,i}}$ values provided for the individual structural groups and interaction factors is an effective predictor of the uncertainty in $\Delta G'^{\circ}_{\mathrm{est}}$. Furthermore, 93% and 99% of the $SE_{\Delta G'^{\circ}_{\mathrm{est}}}$ values for the training set were lower than one and two SE_{MLR}, respectively, verifying that using the $SE_{\Delta G'^{\circ}_{\mathrm{est}}}$ as an estimate of the uncertainty in each $\Delta G'^{\circ}_{\mathrm{est}}$ provides tighter bounds on the uncertainty in the estimates than using the overall SE_{MLR} as the uncertainty estimate for every $\Delta G'^{\circ}_{\mathrm{est}}$.

Results of cross-validation analysis

In addition to assessing the accuracy of the $\Delta_{\mathrm{f}}G'^{\circ}_{\mathrm{est}}$ and $\Delta_{\mathrm{r}}G'^{\circ}_{\mathrm{est}}$ estimates generated by the new group contribution method for the compounds and reactions included in the training set, we also performed a cross-validation analysis to assess the ability of the new method to estimate $\Delta_{\mathrm{f}}G'^{\circ}_{\mathrm{est}}$ and $\Delta_{\mathrm{r}}G'^{\circ}_{\mathrm{est}}$ for compounds and reactions outside the training set. After 200 cross-validation runs were performed (see Methods), the standard error for the data excluded from the training set in the cross-validation runs (SE_{Excluded}) was compared to the standard error for the data remaining in the training set (SE_{MLR}) (Fig. 4 *A*). The overall SE_{Excluded} for all the cross-validation runs was 2.22 kcal/mol, which is only 1.0% higher than the SE_{MLR} for the entire training set (1.90 kcal/mol). These results indicate that the accuracy of $\Delta G'^{\circ}_{\mathrm{est}}$ for the data included in and excluded from the training set is similar. Additionally, the distributions of the residuals for the

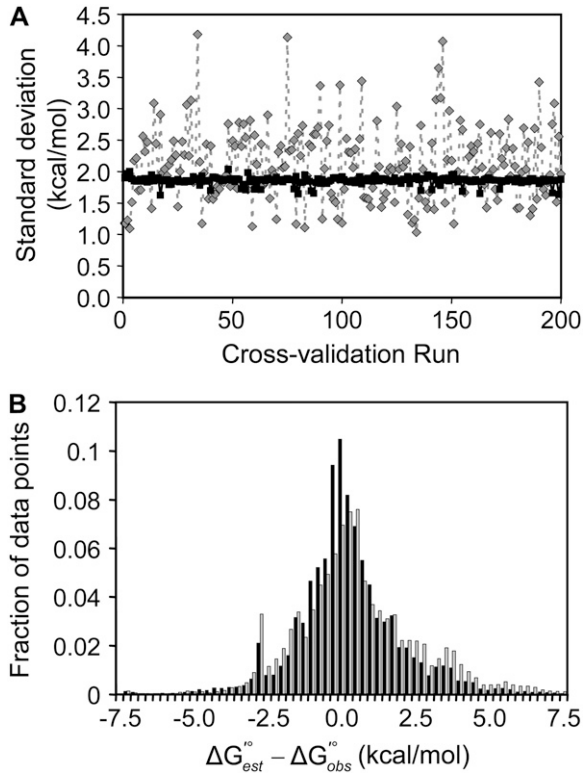


FIGURE 4 Characterization of residuals from the cross-validation analysis. Characterization of residuals for the data associated with the 10% of the reactions and compounds removed from the training set during each cross-validation run. The standard deviation of $\Delta G'^{\circ}_{\mathrm{obs}} - \Delta G'^{\circ}_{\mathrm{est}}$ for the training set (*solid line* and *squares*) varies little over the 200 different samplings performed. The standard deviation of $\Delta G'^{\circ}_{\mathrm{obs}} - \Delta G'^{\circ}_{\mathrm{est}}$ for the data removed from the training set (*gray dashed line* and *diamonds*) varies far more over the 200 different samples performed (*A*). The distribution of all the $\Delta G'^{\circ}_{\mathrm{obs}} - \Delta G'^{\circ}_{\mathrm{est}}$ values for the data removed from the training sets over the 200 cross-validation runs (*shaded bars*) is very similar to the distribution for the data included in the data set, indicating that the accuracy of $\Delta G'^{\circ}_{\mathrm{est}}$ calculated using the group contribution method is similar in magnitude to the accuracy of the fit of the group contribution model to the training set (*B*).

data excluded from the training set (*shaded bars* in Fig. 4 *B*) and the data included in the training set (*solid bars* in Fig. 4 *B*) are nearly identical, further confirming that the accuracy of $\Delta_f G'^{\circ}_{est}$ and $\Delta_r G'^{\circ}_{est}$ for the data included in and excluded from the training set is similar.

To assess the sensitivity of the $\Delta_{gr}G'^{\circ}$ values included in the group contribution method to the training set used to fit the method, we studied the variance of these values during the 200 cross validation runs (Fig. 5). The median $\Delta_{gr}G'^{\circ}$ value calculated for each group during the cross validation analysis never differed from the final reported $\Delta_{gr}G'^{\circ}$ value by more than 0.5 kcal/mol. Furthermore, 50% of the $\Delta_{gr}G'^{\circ}$ values calculated for each group typically fell within 1.0 kcal/mol of the final reported value and always fell within 2 kcal/mol of the final reported value (Fig. 5). These results indicate that the sensitivity of the $\Delta_{gr}G'^{\circ}$ values to the training set used to fit this group contribution method is within the same order of magnitude as the uncertainty in the $\Delta_{gr}G'^{\circ}$ values.

We also examined the accuracy of the $SE_{\Delta G'^{\circ}_{est}}$ values calculated for $\Delta G'^{\circ}_{est}$ of the data excluded from the training set. The residual (difference between $\Delta G'^{\circ}_{est}$ and $\Delta G'^{\circ}_{obs}$) of each excluded data point was compared to the $SE_{\Delta G'^{\circ}_{est}}$ for the same data point, and it was found that 62%, 75%, and 88% of the residuals were less than one $SE_{\Delta G'^{\circ}_{est}}$, two $SE_{\Delta G'^{\circ}_{est}}$, and four $SE_{\Delta G'^{\circ}_{est}}$, respectively. This study indicates that when estimating uncertainty in $\Delta G'^{\circ}_{est}$ for compounds and reactions not included in the data set, uncertainties of two $SE_{\Delta G'^{\circ}_{est}}$ and four $SE_{\Delta G'^{\circ}_{est}}$ will provide approximately the same confidence interval as one and two standard deviations for normally distributed residuals. As a conservative limit, the overall $SE_{Excluded}$ from the cross-validation runs (2.22 kcal/mol) may be used for the uncertainty in any $\Delta G'^{\circ}_{est}$ value, as has been previously proposed by Mavrovouniotis.

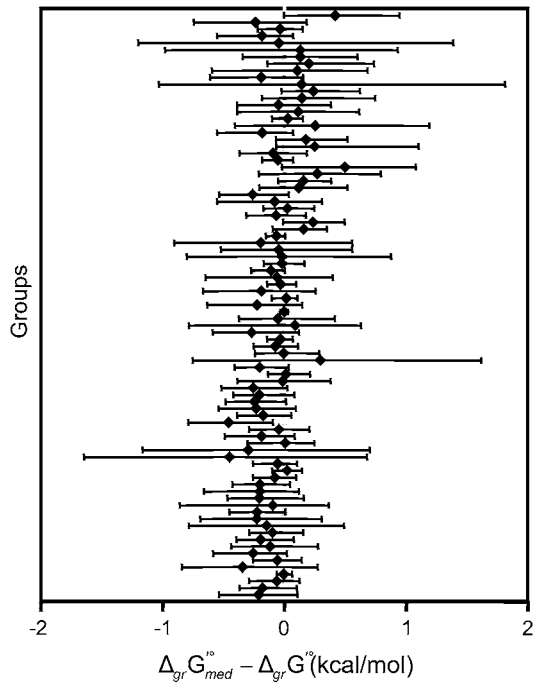


FIGURE 5 Variation of group energy values during cross-validation analysis. The differences between the final reported $\Delta_{gr}G'^{\circ}$ value and the median of the $\Delta_{gr}G'^{\circ}$ values ($\Delta_{gr}G'^{\circ}_{med}$) calculated during the 200 cross-validation runs for each structural group and interaction factor included in the new group contribution method are indicated. The error bars also capture the extent to which the $\Delta_{gr}G'^{\circ}$ value of each group varied from the median value during the cross-validation runs. The error bars left of each point extend through the first quartile of calculated $\Delta_{gr}G'^{\circ}$ values, and the error bars right of each point extend through the third quartile of calculated $\Delta_{gr}G'^{\circ}$ values.

Contribution of the conjugation interaction factors

One significant advance in this new group contribution method compared with previous methods is the addition of interaction factors to account for the effect of double bond conjugation on the $\Delta_f G'^{\circ}$ and $\Delta_r G'^{\circ}$ values. Initially, one new interaction factor was introduced into the group contribution method for each of the types of double bond conjugation possible between carbon, oxygen, and nitrogen atoms (Table 3). Double bond conjugation involving sulfur atoms was not considered, as such structures are less common in biochemistry. Two interaction factors, NCNC and CNNC, were removed from the method before the fitting, as no example of this class of double bond conjugation was found in any of the molecules within the training set. When MLR was used to determine the $\Delta_{gr}G'^{\circ}$, SE_{gr}, and *t*-test values for each of the interaction factors (Table 3), it was found that the interaction factors NCCN and CCNC both had *t*-tests well over 10%, indicating that these interaction factors were not statistically significant. Additionally, the interaction factor OCNC had an insignificant effect on the SE_{MLR} for the fitting. For these reasons, these interaction factors were also removed from the method. All the remaining interaction factors had statistically significant $\Delta_{gr}G'^{\circ}_i$ values, and the addition of each of the remaining interaction factors resulted in a significant drop in the overall SE_{MLR} for the group contribution method. Overall, the inclusion of the five remaining interaction factors for double bond conjugation reduced the SE_{MLR} for the fitting from 2.04 to 1.90 kcal/mol.

Estimating $\Delta G'^{\circ}$ of known biochemical reactions

The group contribution method of Mavrovouniotis and the final implementation of the new group contribution method were both applied to calculating $\Delta_f G'^{\circ}_{est}$ of the compounds and $\Delta_r G'^{\circ}_{est}$ of the reactions in four databases of biochemical reactions: the *i*JR904 genome-scale model of *E. coli*, the *i*AF1260 genome-scale model of *E. coli*, the UM-BBD, and the KEGG (Table 4). The molecular structures of some of the metabolites contained in these databases involve pseudoatoms such as R, X, or *, and the molecular structures of some other metabolites are unknown. These metabolites were considered ineligible for $\Delta_f G'^{\circ}$ estimation because the complete structure of a compound must be known for $\Delta_f G'^{\circ}_{est}$ to be

TABLE 4 Coverage of major biochemical databases

Database	Previous $\Delta_f G'^{\circ}_{est}$*	New $\Delta_f G'^{\circ}_{est}$	Eligible metabolites†	Previous $\Delta_r G'^{\circ}_{est}$*	New $\Delta_r G'^{\circ}_{est}$	Eligible reactions‡
*i*JR904	543 (91%)	570 (96%)	595	864 (93%)	905 (97%)	931
*i*AF1260	837 (90%)	872 (94%)	925	1933 (93%)	1996 (96%)	2077
UM-BBD	667 (66%)	811 (80%)§	1013	548 (56%)	777 (79%)§	983
KEGG§	6429 (50%)	8186 (64%)	12,759	4542 (84%)	4945 (93%)	5402

*Previous $\Delta_f G'^{\circ}_{est}$ and $\Delta_r G'^{\circ}_{est}$ estimates were generated using the group contribution method of Mavrovouniotis (9,10).
†Only metabolites with known molecular structures that do not involve any pseudoatoms like R, X, or * are eligible for $\Delta_f G'^{\circ}_{est}$ calculation.
‡Only mass and charge balanced reactions that do not involve any metabolites with unknown molecular structures are eligible for $\Delta_r G'^{\circ}_{est}$ calculation. §These estimates were initially reported and discussed in detail in an earlier work (7).
§1/13/2008 build of the KEGG.

calculated. Similarly, some of the reactions contained in these databases are not mass or charge balanced or involve compounds with unknown molecular structures. Such reactions were also considered ineligible for $\Delta_r G'^{\circ}$ estimation because $\Delta_r G'^{\circ}_{est}$ can be calculated only for complete and balanced reactions. Once the ineligible reactions and compounds were removed from consideration, any remaining compounds and reactions for which $\Delta_f G'^{\circ}_{est}$ and $\Delta_r G'^{\circ}_{est}$ could not be calculated were entirely due to the presence of molecular substructures for which the $\Delta_{gr} G'^{\circ}_i$ value was unknown.

Both the group contribution method of Mavrovounoitis and the new group contribution method are capable of estimating $\Delta_f G'^{\circ}$ and $\Delta_r G'^{\circ}$ for nearly all the compounds and reactions involved in the *i*JR904 and *i*AF1260 models. The coverage of the new group contribution method is only slightly better for these genome-scale models. However, the new group contribution method performs significantly better than the Mavrovouniotis method in estimating $\Delta_f G'^{\circ}$ and $\Delta_r G'^{\circ}$ of the UM-BBD compounds and reactions. This is primarily due to the addition of $\Delta_{gr} G'^{\circ}_i$ values for the halogen substructures, which are prevalent in the biodegradation chemistry. The new group contribution method also performs significantly better in estimating $\Delta_f G'^{\circ}$ and $\Delta_r G'^{\circ}$ of the KEGG compounds and reactions. All 20 additional substructures that have been included in the new group contribution method contribute evenly to this improvement in the coverage of the KEGG. All the $\Delta_f G'^{\circ}$ and $\Delta_r G'^{\circ}$ values estimated for the compounds and reactions in the KEGG using the new group contribution method have been provided in Data S2. Note that in all four databases, the coverage of the reactions by the group contribution method is better than the coverage of the compounds. This is because structural groups with unknown $\Delta_{gr} G'^{\circ}_i$ values cancel out of many reactions, as they are not created or destroyed in most reactions. Overall, the new group contribution method is demonstrated to be capable of estimating $\Delta_f G'^{\circ}$ and $\Delta_r G'^{\circ}$ for a wide range of biochemical compounds and reactions.

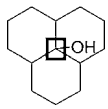
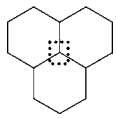
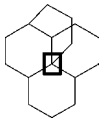
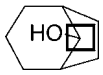
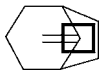
Prevalent substructures with unknown $\Delta_{gr} G'^{\circ}$ values

Clearly, molecular substructures still exist in these databases for which the $\Delta_{gr} G'^{\circ}_i$ value is unknown. Many of these structures are present in organic-inorganic complexes involving iron, nickel, or cobalt for which the new group contribution method has not been designed. However, a small number of prevalent organic substructures with unknown $\Delta_{gr} G'^{\circ}$ values appear in many of the metabolites and reactions for which $\Delta_f G'^{\circ}_{est}$ and $\Delta_r G'^{\circ}_{est}$ could not be calculated (Table 5). These substructures represent important targets for future experiments involving the measurement of the thermodynamic properties of biochemical reactions. As experimental data for reactions involving these substructures does emerge, new structural groups and interaction factors can be developed and added to this group contribution method. When determining $\Delta_{gr} G'^{\circ}$ values for these additions to the model, it is recommended that the new $\Delta G'^{\circ}_{obs}$ data be appended to the training set used to fit the entire group contribution model and that all $\Delta_{gr} G'^{\circ}$ values be refit in the model rather than solely the values for the new groups. This will result in better accuracy and reveal the effect of the addition of the new data and groups on the method. To facilitate this kind of expansion and improvement of this group contribution method, details of the training set used in this method have been provided in Data S2. Molfiles created for the molecular structures of every compound involved in the training set in the correct ionic form at pH 7 are also available in Data S3.

DISCUSSION

The group contribution method introduced here has numerous advantages over previous methods including i), the ability to calculate $\Delta G'^{\circ}_{est}$ for a greater variety of compounds and reactions; ii), improved accuracy in the $\Delta G'^{\circ}_{est}$ values produced using the method; iii), improved estimation for the uncertainty in the $\Delta G'^{\circ}_{est}$ values produced; and iv), complete disclosure of the training set used to fit the $\Delta_{gr} G'^{\circ}$ values to facilitate the expansion of the method with additional data, interaction factors, and structural groups. The application of this group contribution method toward the estimation of $\Delta_f G'^{\circ}_{est}$ and $\Delta_r G'^{\circ}_{est}$ for the compounds and reactions in the *i*JR904 model, the *i*AF1260 model, the UM-BBD, and the KEGG confirms the ability of the method to predict $\Delta_f G'^{\circ}_{est}$ and $\Delta_r G'^{\circ}_{est}$ for a significant portion of the known

TABLE 5 Prevalent molecular substructures with unknown $\Delta_{gr}G_i'^{\circ}$ values

Substructure description	Example structure	KEGG compounds involving substructure
>C< (in three fused rings)		410
>CH- (in three fused rings)		390
$>CH_2$ (in two fused rings)		255
$-NO_2$		162
>C< (in four fused rings)		143
>C= (in three fused rings, one aromatic)		139
>CH- (in two fused rings)		127
>C= (in two fused rings)		71

biochemistry. The $\Delta_f G'^{\circ}_{est}$ and $\Delta_r G'^{\circ}_{est}$ estimations generated for the KEGG and provided in Data S2 represent the most complete and most accurate set of thermodynamic data compiled for the KEGG to date, to our knowledge, and the addition of halogens to the methodology allows the application of this method to new types of chemistry beyond genome-scale metabolic models in areas such as bioremediation (24).

All the changes introduced in this new group contribution method have not only expanded the applicability of the method to calculate $\Delta G'^{\circ}_{est}$ for a wider range of compounds and reactions but also improved the accuracy of the method. For the compounds and reactions in the training set for which $\Delta G'^{\circ}_{est}$ can be calculated using the Mavrovouniotis group contribution method, the standard deviation of the residuals is 3.92 kcal/mol, compared to a standard deviation of 1.98 kcal/mol when the new group contribution method is used to calculate $\Delta G'^{\circ}_{est}$ for the same reactions and compounds.

The quantification of the uncertainty in each $\Delta_{gr}G_i'^{\circ}$ value in the method allows improved resolution in uncertainty estimates for all $\Delta G'^{\circ}_{est}$ produced using the method. This enhanced resolution in the uncertainty in $\Delta G'^{\circ}_{est}$ is essential to any genome-scale analysis of metabolic pathways and metabolomic studies involving thermodynamics such as thermodynamics-based metabolic flux analysis (2). As a result of the cross-validation analysis, it is recommended that the uncertainty used for $\Delta G'^{\circ}_{est}$ values calculated with this method be four times the $SE_{\Delta G'^{\circ}_{est}}$ calculated from the SE_{gr} values using Eq. 10, as this uncertainty provides an 83% confidence interval for $\Delta G'^{\circ}_{est}$.

A web interface has been developed to allow the automated estimation of the $\Delta_f G'^{\circ}$ values for a set of compounds based on the molecular structures of the compounds using the new group contribution method. This interface is available free at the following web address: http://sparta.chem-eng.northwestern.edu/cgi-bin/GCM/WebGCM.cgi.

SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

Many thanks to Stacey Pace for her assistance with the UM-BBD database analysis.

This work was supported by the U.S. Department of Energy Genomes to Life program, the DuPont Young Professor's grant, and a National Science Foundation Integrative Graduate Education and Research Traineeship Complex Systems fellowship.

REFERENCES

1. Henry, C. S., M. D. Jankowski, L. J. Broadbelt, and V. Hatzimanikatis. 2006. Genome-scale thermodynamic analysis of Escherichia coli metabolism. *Biophys. J.* 90:1453–1461.
2. Henry, C. S., L. J. Broadbelt, and V. Hatzimanikatis. 2007. Thermodynamics-based metabolic flux analysis. *Biophys. J.* 92:1792–1805.
3. Price, N. D., J. A. Papin, C. H. Schilling, and B. O. Palsson. 2003. Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol.* 21:162–169.
4. Feist, A. M., C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1261 ORFs and thermodynamic information. *Mol. Syst. Biol.* 3:1–18.
5. Kummel, A., S. Panke, and M. Heinemann. 2006. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.* 2:1–10.
6. Kummel, A., S. Panke, and M. Heinemann. 2006. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics.* 7:1–12.
7. Reference deleted in proof.
8. Beard, D. A., and H. Qian. 2005. Thermodynamic-based computational profiling of cellular regulatory control in hepatocyte metabolism. *Am. J. Physiol. Endocrinol. Metab.* 288:E633–E644.
9. Mavrovouniotis, M. L. 1991. Estimation of standard Gibbs energy changes of biotransformations. *J. Biol. Chem.* 266:14440–14445.
10. Mavrovouniotis, M. L. 1990. Group contributions for estimating standard Gibbs energies of formation of biochemical-compounds in aqueous-solution. *Biotechnol. Bioeng.* 36:1070–1082.
11. Benson, S. W. 1968. Thermochemical Kinetics; Methods for the Estimation of Thermochemical Data and Rate Parameters. Wiley, New York.
12. Maskow, T., and U. V. Stockar. 2005. How reliable are thermodynamic feasibility statements in biochemical pathways? *Biotechnol. Bioeng.* 92:223–230.

13. Scholten, J. C. M., J. C. Murrell, and D. P. Kelly. 2003. Growth of sulfate-reducing bacteria and methanogenic archaea with methylated sulfur compounds: a commentary on the thermodynamic aspects. *Arch. Microbiol.* 179:135–144.

14. VanBriesen, J. M. 2002. Evaluation of methods to predict bacterial yield using thermodynamics. *Biodegradation.* 13:171–190.

15. Weber, A. L. 2002. Chemical constraints governing the origin of metabolism: the thermodynamic landscape of carbon group transformations under mild aqueous conditions. *Orig. Life Evol. Biosph.* 32:333–357.

16. Magnus, J. B., D. Hollwedel, M. Oldiges, and R. Takors. 2006. Monitoring and modeling of the reaction dynamics in the valine/leucine synthesis pathway in Corynebacterium glutamicum. *Biotechnol. Prog.* 22:1071–1083.

17. Hatzimanikatis, V., C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, and L. J. Broadbelt. 2004. Exploring the diversity of complex metabolic networks. *Bioinformatics.* 21:1603–1609.

18. Li, C., J. A. Ionita, C. S. Henry, M. D. Jankowski, V. Hatzimanikatis, and L. J. Broadbelt. 2004. Computational discovery of biochemical routes to specialty chemicals. *Chem. Eng. Sci.* 59:5051–5060.

19. Tanaka, M., Y. Okuno, T. Yamada, S. Goto, S. Uemura, and M. Kanehisa. 2003. Extraction of a thermodynamic property for biochemical reactions in the metabolic pathway. *Genome Inform.* 14: 370–371.

20. Caspi, R., H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S. Y. Rhee, C. Tissier, P. F. Zhang, and P. D. Karp. 2006. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 34:D511–D516.

21. Schomburg, I., A. J. Chang, O. Hofmann, C. Ebeling, F. Ehrentreich, and D. Schomburg. 2002. BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem. Sci.* 27:54–56.

22. Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27:29–34.

23. Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res.* 30:42–46.

24. Ellis, L. B. M., D. Roe, and L. P. Wackett. 2006. The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res.* 34:D517–D521.

25. Alberty, R. A. 1998. Calculation of standard transformed formation properties of biochemical reactants and standard apparent reduction potentials of half reactions. *Arch. Biochem. Biophys.* 358:25–39.

26. Alberty, R. A. 2003. Thermodynamics of Biochemical Reactions. Massachusetts Institute of Technology Press, Cambridge, MA.

27. Goldberg, R. N., Y. B. Tewari, and T. N. Bhat. 2004. Thermodynamics of enzyme-catalyzed reactions—a database for quantitative biochemistry. *Bioinformatics.* 20:2874–2877.

28. Thauer, R. K., K. Jungermann, and K. Decker. 1977. Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol. Rev.* 41:100–180.

29. Thauer, R. K. 1998. Biochemistry of methanogenesis: a tribute to Marjory Stephenson. *Microbiology.* 144:2377–2406.

30. Dolfing, J., and B. K. Harrison. 1992. Gibbs free-energy of formation of halogenated aromatic-compounds and their potential role as electron-acceptors in anaerobic environments. *Environ. Sci. Technol.* 26:2213–2218.

31. Dolfing, J., and D. B. Janssen. 1994. Estimates of Gibbs free energies of formation of chlorinated aliphatic compounds. *Biodegradation.* 5:21–28.

32. Forsythe, R. G., P. D. Karp, and M. L. Mavrovouniotis. 1997. Estimation of equilibrium constants using automated group contribution methods. *Comput. Appl. Biosci.* 13:537–543.

33. Wade, L. G. 2003. Organic Chemistry. Prentice Hall/Pearson Education, Upper Saddle River, NJ.

34. Neter, J., W. Wasserman, and M. H. Kutner. 1990. Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. Irwin, Homewood, IL.

35. Wagman, D. D. 1982. The NBS Tables of Chemical Thermodynamic Properties: Selected Values for Inorganic and C1 and C2 Organic Substances in SI Units. American Chemical Society and the American Institute of Physics for the National Bureau of Standards, Washington, DC.

36. Alberty, R. A. 1998. Calculation of standard transformed Gibbs energies and standard transformed enthalpies of biochemical reactants. *Arch. Biochem. Biophys.* 353:116–130.

37. Soovali, L., E. I. Room, A. Kutt, I. Kaljurand, and I. Leito. 2006. Uncertainty sources in UV-Vis spectrophotometric measurement. *Accredit. Qual. Assur.* 11:246–255.

38. Reed, J. L., T. D. Vo, C. H. Schilling, and B. O. Palsson. 2003. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol.* 4:1–12.